This document describes how to use the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus command line interface of the GenBank Submission Portal. This tool can be used for large and/or frequent submissions of SARS-CoV-2 complete or partial genomes and is for submitting only SARS-CoV-2 sequences.

Users should contact gb-admin@ncbi.nlm.nih.gov to discuss requirements for submission through the programmatic interface. An FTP upload directory will be created for each group to deliver their submission files. Submitters must also have a MyNCBI user account (https://www.ncbi.nlm.nih.gov/account/register/).

The following information will be required to establish an account:

- MyNCBI account email of the primary submitter
- center/account abbreviation
- full center/account name
- names and email addresses of all additional users
- postal address of institute (including postal code and country)

For each new submission, create a new submission folder under the FTP upload directory. The submission folder must have group read, write, and execute permissions (use Linux commands "ls -dl" to see permissions, and "chmod 775 <folder name>" to change them if needed). The FTP account will be provided with "Production" and "Test" directories. Submission folders under the "Test" directory can be used to familiarize yourself with the submission process. Final submission folders should be created under the "Production" directory.

Data Files

A .zip archive file containing the following should be uploaded to the appropriate submission folder. Each file must have a specific extension, as shown in parentheses below. Only .zip archives can be used at this time. Annotation files should **not** be included. The submission will be automatically annotated by the submission pipeline. Any additional files beyond the files listed below will cause the submission to fail.

Description of the files and mandatory file extensions:

- **Sequence data (.fsa)** Nucleotide sequences in FASTA format.
    - sequence IDs must be unique for each sequence
    - cannot contain spaces
    - may contain only the following characters - letters, digits, hyphens (-), underscores (_), periods (.), colons (:), asterisks (*), and number signs(#).
    - should be under 25 characters

  For more information, see: https://submit.ncbi.nlm.nih.gov/genbank/help/#fasta
    - **(OPTIONAL, for SPHERES submitters only)** purpose of sequencing tracking keyword: Your FASTA file should contain the following in the FASTA definition line, separated from the Sequence ID by a space *[keyword=purposeofsampling:baselinesurveillance]*. Note this tag should appear in each FASTA definition line.

```
>Seq1 [keyword=purposeofsampling:baselinesurveillance]
CTAGCTAGCTAGCTAGCTAGCTAGCTAG
>Seq2 [keyword=purposeofsampling:baselinesurveillance]
CTAGCTAGCTAGCTAGCTAGCTAGCTAG
```

- **Source metadata table (.src)** Tab-delimited text file which must include: sequence_ID, full organism name, isolate, collection-date, host, country, isolation-source (optional). Other source metadata may also be included; see here for a list of all source modifiers: https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers.

- sequence_ID: must match the sequence_ID in the FASTA sequence data file
- organism: must be *Severe acute respiratory syndrome coronavirus 2*
- isolate: ICTV formatted isolate name
    - SARS-CoV-2/host/three letter country abbreviation/sample ID/year
      `SARS-CoV-2/human/USA/GA_2741/2020`
        - host: common or scientific name of the host animal from which the virus was located, for example: Homo sapiens
- collection-date: date the sample was collected in ISO format, for example: 2020-03-30
- country: the country where the sample was isolated.
  See https://www.ncbi.nlm.nih.gov/genbank/collab/country/ for INSDC country list. Additional locality information can be included after the colon in the country, for example: USA: Maryland
- isolation-source: (OPTIONAL) the physical source of the viral sample, for example: nasal swab
- BioProject, BioSample, SRA numbers (OPTIONAL)

For more information on formatting a .src table, see:
https://submit.ncbi.nlm.nih.gov/genbank/help/#srcmods

- **Submission template (.sbt)** Text file with submitter names and organizations, as well as publications associated with or describing the sequence. Users can generate submission template files at https://submit.ncbi.nlm.nih.gov/genbank/template/submission/. The saved template can be reused for multiple submissions.
- **Structured comment (.cmt)** This is an OPTIONAL tab-delimited text file which can be used to provide additional sequencing technology and assembly information. For more information, please see https://www.ncbi.nlm.nih.gov/genbank/structuredcomment/#GenBank

See the example.zip file located with this document for an example data file archive and file formats.

submission.xml File

In addition to the archive file, users must upload a separate submission form with the file name "submission.xml". submission.xml file acts like an envelope to the submission and includes the necessary instructions on how to process this submission. The schema for the submission.xml is defined in submission.xsd: http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/submit/public-docs/common/. The submission.xml must **not** be included in the data file .zip archive.

Sample submission.xml file:

```
<?xml version="1.0"?>
<Submission>
  <Description>
    <Title>Submission title, ASCII characters only, 512 characters max</Title>
    <Comment>SARS-CoV-2 test submission</Comment>
    <Organization type="center" role="owner">
      <Name>account name</Name>
    </Organization>
    <Hold release_date="2024-05-25"/>
  </Description>
  <Action>
    <AddFiles target_db="GenBank">
      <File file_path="sarscov2.zip">
        <DataType>genbank-submission-package</DataType>
      </File>
      <Attribute name="wizard">BankIt_SARSCoV2_api</Attribute>
      <Attribute name="auto_remove_failed_seqs">no</Attribute>
      <Identifier>
        <SPUID spuid_namespace="ncbi-sarscov2-genbank">2020-03-
04.sarscov2</SPUID>
      </Identifier>
    </AddFiles>
  </Action>
</Submission>
```

- The `Description` block contains user information and submission comments. `Comment` is an optional line for user tracking. The `Organization` line is required with `type="center" role="owner"`. The **OPTIONAL** `Title` line overrides the default submission name in the Submission Portal. This field can be used to customize submission names for internal tracking. The `Name` field is for the user account name that will be provided when the submission account is established.

```
<Comment>SARS-CoV-2 test submission</Comment>
<Organization type="center" role="owner">
      <Name>account name</Name>
```

- If the data has a specific release date, this should be included in the `Hold release_date`

```
<Hold release_date="2028-05-25"/>
```

  - If this line is omitted, the data will be released immediately after processing.
- The `AddFiles` block contains required information for specifying the data file archive and directing the submission to the correct pipeline. The `target_db` and **`wizard`** fields are required lines that cannot be changed. The data file .zip file name must match the `file_path` in the submission.xml file:

```
<AddFiles target_db="GenBank">
   <File file_path="sarscov2.zip">
         <DataType>genbank-submission-package</DataType>
   </File>
   <Attribute name="wizard">BankIt_SARSCoV2_api</Attribute>
   <Attribute name="auto_remove_failed_seqs">no</Attribute>
```

  - The **OPTIONAL** **`auto_remove_failed_seqs`** line instructs the pipeline how to handle sequences with errors. **`auto_remove_failed_seqs`** only applies to submissions with more than one sequence. The parameter is ignored for single sequence submissions. The only valid inputs for this field are '`yes`' or '`no`'.
    - `no`: sequences with errors are reported back to the submitter and the entire submission is held from further processing until reviewed and corrected by the submitter.

  - `yes`: sequences with errors are automatically removed from the submission. The remaining sequences that pass annotation and validation are processed. A list of removed sequences and the reason for removal are reported back to the user in the Submission Portal.

- SPUID (Submitter Provided Unique Identifier) is a user-generated identification number. Before an accession number is assigned, the SPUID allows submitters to keep track of samples as they are being processed by NCBI. The SPUID will not appear in the final record and should not be included in a publication.

```
<SPUID spuid_namespace="ncbi-sarscov2-genbank">2020-03-
04.sarscov2</SPUID>
```

  - The value for `spuid_namespace` in the first part of the line (**`ncbi-sarscov2-genbank`**) is the center/account abbreviation provided during account creation. This value will remain the same for every submission.
  - The **second part** of the line (**`2020-03-04.sarscov2`**) **must be unique** for each submission from the submitter.

Submitting the Data

When the data files and submission.xml are ready to be submitted, upload an empty text file named "submit.ready" into the submission folder. The submit.ready file must have user and group read and write permissions (chmod 664 <file name>).

For example, a submission might look like this:
FTP upload directory (*This is an example and will not work. Please contact gb-admin@nlm.nih.gov to have an FTP upload directory generated for you or your group*):
`ftp://login@ftp-private.ncbi.nlm.nih.gov/`

Submission folder:
`ftp://login@ftp-private.ncbi.nlm.nih.gov/Production/20200420/`

Upload these files to folder 20200420:

1. .fsa, .sbt, .src, (optional .cmt) – all together in "**`sarscov2.zip`**"
2. `submission.xml` – references **`sarscov2.zip`** in the `file_path` line
3. `submit.ready`

`The submission folder must contain only the data.zip file, the submission.xml file and the submit.ready file. Any additional files or subfolders will cause the submission to fail.`

After submission

- The Submission Portal software scans the upload directories several times per day. When it finds a new "submit.ready" file, it checks the submission.xml file and looks for the associated archive file. If all the required data files are present, processing begins.
- Submission Portal provides diagnostic message with listing of failed actions and files and provides error descriptions. Submission Portal creates submission report file(s) in the submission folder. The submission report has the name "report.<N>.xml", where <N> stands for consecutive numbers 1, 2, etc. The first report file made by Submission Portal is always "report.1.xml". Additional updates will be reported in "report.2.xml," "report.3.xml," etc.

- Errors such as the absence of required files, presence of more files than referenced in submission.xml or file formatting issues, will stop the submission before a submission number is assigned. If there are no errors, the report.xml will have the submission (SUB) number and the status of the submission. When a submission is successfully received by the pipeline, the original submission data files will be **removed** from that folder. This is to avoid duplicate submissions. If errors are reported, the original submission data file is **retained** in the directory. Correct the data files and generate a new submit.ready file to initiate a resubmission.
    - Note: a submission (SUB) number is not an accession number. It will not appear in the final record and should not be cited in a publication.
- Submissions that are successfully received for processing will undergo automated sequence checks:
    - **trim** terminal NNNs and ambiguous sequence ends
    - **remove** low quality sequences with >50% ambiguous nucleotides
    - **remove** sequences that are entirely vector
    - **remove** sequences below the minimal sequence length: under 50 nt.
    - **remove** sequences longer than the expected length: 30,000 nt. for SARS-CoV-2
- Submitted sequences will be automatically annotated by the Viral Annotation DefineR tool (VADR). We strongly encourage you to check your sequences using VADR before submission. Correct any errors before submitting. Error codes are described in this table. For an example of an annotated genome, see: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512
    - VADR: https://github.com/nawrockie/vadr/blob/master/README.md
    - VADR SARS-CoV-2 annotation: https://github.com/nawrockie/vadr/wiki/Coronavirus-annotation
- Submissions that are processed by the pipeline will be reviewed by GenBank curation staff. Submissions with annotation errors will be returned to the submitter. Submissions that pass curation will be assigned accession numbers. The accession numbers, copies of the annotated flatfiles, and summary files are reported on the submission portal under the SUB number: https://submit.ncbi.nlm.nih.gov/subs/api/
- Reports can also be obtained programmatically.
    - In the 'report.xml', look at 'File' XML elements and get their 'file_id' attribute value.
    - Use wget, curl or any equivalent tool to retrieve file from this URL:
    - curl 'https://submit.ncbi.nlm.nih.gov/api/2.0/files/<@file_id@>/?format=attachment'
    - Documentation on the API is here: https://submit.ncbi.nlm.nih.gov/api/2.0/docs/